Supplementary Material

S.1 The Sellke ZigZag process

In this section we describe a method for augmenting the state space to relate the NuZZ to the construction [Sellke, 1983] discussed in Section 3.1. This can be done by changing the structure of the Zig-Zag sampler, from a PDMP that jumps in conjunction with an arrival from a Poisson process, to a PDMP that jumps when one of the components touches an active boundary [Davis, 1993]. We call this Markov process the Sellke Zig-Zag (SeZZ), and this can be shown using methods from Davis [1993] to target the correct invariant measure. All the processes considered in this section are one-dimensional ($d = 1$).

Let us define a discrepancy variable $q(t)$ as

$$q(t) = R - \int_0^t \Lambda(x(s), v)\mathrm{d}s\,, \qquad t \in [0, \tau]\,. \tag{S.1}$$

The evolution of $q(t)$ in time can be seen as a jump process where the value of $q(t)$ decreases deterministically according to

$$\frac{\mathrm{d}q}{\mathrm{d}t} = -\Lambda(x(t), v) \tag{S.2}$$

until it reaches $q(\tau^-) = 0$. At that point it jumps to $q(\tau^+) = R \sim \mathrm{Exp}(1)$, and continues decreasing according to Equation (S.2), analogously to the way in which the differential equation (1) determines the dynamics of the state x. By adding the discrepancy variable $q(t)$ to the existing position $x(t)$ and velocity $v(t)$ variables, we can now define the SeZZ process on (x, v, q) as a PDMP where a velocity flip is triggered by the variable $q(t)$ hitting the active boundary at 0.

In order to prove that this process targets the correct stationary distribution, we will prove that the forward Kolmogorov equation is equal to zero. However, as this process has an active boundary in the state space, the generator has to differ from Equation (3) using methods from [Davis, 1993, p.118].

Let the variable q be defined on the space $\mathcal{Q} = [0, \infty)$, which can be partitioned into the subsets $\mathcal{Q}_0 = (0, \infty)$ and $\{0\}$. Let $E = \mathcal{X} \times \mathcal{V} \times \mathcal{Q}_0$ be the interior of the state space, with $\mathcal{X} = \mathbb{R}$ and $\mathcal{V} = \{-1, +1\}$, and $\mathcal{B} = \mathcal{X} \times \mathcal{V} \times \{0\}$ be an active boundary of the state space. While travelling through E, the process is described by the drift operator

$$\mathcal{D}f(x, v, q) = v\frac{\partial f}{\partial x} - \Lambda(x, v)\frac{\partial f}{\partial q}\,, \tag{S.3}$$

meaning that motion is completely deterministic and there are no jumps. When the variable $q(t)$ reaches zero, the process touches the active boundary $\mathcal{B}$ causing the velocity to switch, and q is refreshed with a new exponential random variable. This change is expressed through the boundary operator

$$\mathcal{C}f(x, v, 0) = \int_{\mathcal{Q}_0} \mathrm{e}^{-q} f(x, -v, q)\mathrm{d}q - f(x, v, 0)\,. \tag{S.4}$$

Let us define the stationary measure μ as

$$\mu(\mathrm{d}x \times \mathrm{d}v \times \mathrm{d}q) = \pi(x)\,\psi(v)\,\varphi(q)\,\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}q\,, \tag{S.5}$$

where $\pi(x)$ is the stationary distribution of interest, $\psi(v)$ is the stationary distribution of the velocities, i.e. $1/2$ on both -1 and 1, and $\psi(q)$ is the stationary distribution of the discrepancy variable q, which we will show to be e^{-q} in our case.

The process SeZZ targets the stationary measure μ defined above in Equation (S.5) if

$$\int_E \mathcal{D}f(x, v, q)\,\mu(\mathrm{d}x \times \mathrm{d}v \times \mathrm{d}q) + \int_{\mathcal{B}} \mathcal{C}f(x, v, 0)\,\rho(\mathrm{d}x \times \mathrm{d}v) = 0\,, \tag{S.6}$$

where ρ is called the boundary measure, and f is a real, absolutely continuous function whose jumps at the boundary are finite [Davis, 1993, p.118]. See the proof below for a more precise description.

Theorem 2. *The one-dimensional SeZZ process with extended generator given by the operators*

$$\mathcal{D}f(x,v,q) = v\frac{\partial f}{\partial x} - \Lambda(x,v)\frac{\partial f}{\partial q} \tag{S.7}$$

and

$$\mathcal{C}f(x,v,0) = \int_{\mathcal{Q}_0} \mathrm{e}^{-q} f(x,-v,q)\mathrm{d}q - f(x,v,0) \tag{S.8}$$

targets has the joint measure μ as its invariant measure, which admits π as a marginal.

Proof. Let the function $f : E \to \mathbb{R}$ be absolutely continuous and measurable [Davis, 1993, p.118, 82], and let $\int_E |f(x,v,q) - f(x_{t^-}, v_{t^-}, q_{t^-})|\mathrm{d}\varsigma < \infty$, where ς is the measure $\varsigma = \sum_{T_k \le t} \delta((x_{T_k}, v_{T_k}, q_{T_k}), T_k)$, $\forall t \ge 0$, with T_k representing the jumping times and δ_r being the Dirac measure at $r \in E \times \mathbb{R}$ [Davis, 1984, p.367]. Let the stationary measure be as in Equation (S.5), and following Löpker and Palmowski [2013], let the boundary measure be

$$\rho\,(\mathrm{d}x \times \mathrm{d}v) = \Lambda(x,v)\pi(x)\,\psi(v)\,\varphi(0)\,\mathrm{d}x\,\mathrm{d}v\,. \tag{S.9}$$

Following [Davis, 1993, p.118], we will show that SeZZ has $\pi(x)$ as marginal stationary distribution by proving that Equation (S.6) is satisfied for

$$\pi(x)\,\psi(v)\,\varphi(q) = \frac{1}{Z}\mathrm{e}^{-U(x)} \times \frac{1}{2} \times \mathrm{e}^{-q}\,. \tag{S.10}$$

Substituting the quantities above into Equation (S.6) above, and assuming that $\pi(x) = 0$ at $x = \pm\infty$, we obtain

$$\begin{aligned}\int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \left(v\,\frac{\partial f}{\partial x}\bigg|_{(x,v,q)} - \Lambda(x,v)\,\frac{\partial f}{\partial q}\bigg|_{(x,v,q)} \right) \mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x \\ + \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \left(\int_{\mathcal{Q}_0} f(x,-v,q)\mathrm{e}^{-q}\mathrm{d}q - f(x,v,0) \right) \Lambda(x,v)\,\mathrm{e}^{-U(x)}\,\mathrm{d}x = 0\,,\end{aligned} \tag{S.11}$$

where the common constant $1/(2Z)$ has been cancelled. The equality follows by integration by parts and rearranging terms.

Our starting point is Equation (S.11) above, which we will re-write as:

$$\begin{aligned}&\overbrace{\int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} v\,\frac{\partial f}{\partial x}\bigg|_{(x,v,q)} \mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x}^{\text{First term, } I_1} + \\ &\overbrace{-\int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \Lambda(x,v)\,\frac{\partial f}{\partial q}\bigg|_{(x,v,q)} \mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x}^{\text{Second term, } I_2} \\ &\underbrace{+\int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \left(\int_{\mathcal{Q}_0} f(x,-v,q)\mathrm{e}^{-q}\mathrm{d}q - f(x,v,0) \right) \Lambda(x,v)\,\mathrm{e}^{-U(x)}\,\mathrm{d}x}_{\text{Third term, } I_3} = 0\,.\end{aligned} \tag{S.12}$$

To verify this equation, we proceed by analysing each of the three terms individually. We will often exchange the order of integration in integrals, making use of the Fubini-Tonelli theorem. Rearranging the factors and applying integration by parts, the first term becomes

$$\begin{aligned} I_1 &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} v\,\frac{\partial f}{\partial x}\bigg|_{(x,v,q)} \mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x \\ &= \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} v\mathrm{e}^{-q} \int_{\mathcal{X}} \frac{\partial f}{\partial x}\bigg|_{(x,v,q)} \mathrm{e}^{-U(x)}\,\mathrm{d}x\,\mathrm{d}q \\ &= \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} v\mathrm{e}^{-q} \left([f(x,v,q)\mathrm{e}^{-U(x)}]_{x=-\infty}^{\infty} + \int_{\mathcal{X}} f(x,v,q)\frac{\mathrm{d}U}{\mathrm{d}x}\mathrm{e}^{-U(x)}\,\mathrm{d}x \right) \mathrm{d}q \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} v f(x,v,q)\frac{\mathrm{d}U}{\mathrm{d}x}\mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x\,, \end{aligned} \tag{S.13}$$

where the term in square brackets on the third line is zero as we work with target densities $\pi(x) \propto \mathrm{e}^{-U(x)}$ that tend to zero as $x \to \pm\infty$. Again, we rearrange the factors and apply integration by parts to the second term:

$$\begin{aligned} I_2 &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \Lambda(x,v) \left.\frac{\partial f}{\partial q}\right|_{(x,v,q)} \mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \Lambda(x,v)\mathrm{e}^{-U(x)} \int_{\mathcal{Q}_0} \left.\frac{\partial f}{\partial q}\right|_{(x,v,q)} \mathrm{e}^{-q}\,\mathrm{d}q\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \Lambda(x,v)\mathrm{e}^{-U(x)} \left([f(x,v,q)\mathrm{e}^{-q}]_{q=0}^{\infty} - \int_{\mathcal{Q}_0} f(x,v,q)\frac{\mathrm{d}\mathrm{e}^{-q}}{\mathrm{d}q}\,\mathrm{d}q \right) \mathrm{d}x \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \Lambda(x,v) f(x,v,q)\mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x - \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \Lambda(x,v) f(x,v,0)\mathrm{e}^{-U(x)}\,\mathrm{d}x\,. \end{aligned} \tag{S.14}$$

Finally, the third term reduces to

$$\begin{aligned} I_3 &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \left(\int_{\mathcal{Q}_0} f(x,-v,q)\mathrm{e}^{-q}\,\mathrm{d}q - f(x,v,0) \right) \Lambda(x,v)\mathrm{e}^{-U(x)}\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} f(x,-v,q)\mathrm{e}^{-q}\Lambda(x,v)\mathrm{e}^{-U(x)}\,\mathrm{d}q\,\mathrm{d}x - \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} f(x,v,0)\Lambda(x,v)\mathrm{e}^{-U(x)}\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \Lambda(x,-v) f(x,v,q)\mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x - \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \Lambda(x,v) f(x,v,0)\mathrm{e}^{-U(x)}\,\mathrm{d}x\,. \end{aligned} \tag{S.15}$$

Substituting (S.13), (S.14) and (S.15) into (S.12), we obtain

$$I_1 - I_2 + I_3 = \int_{\mathcal{X}} \sum_{v\in\mathcal{V}} \int_{\mathcal{Q}_0} \left(v\frac{\mathrm{d}U}{\mathrm{d}x} - (\Lambda(x,v) - \Lambda(x,-v)) \right) f(x,v,0)\,\mathrm{e}^{-U(x)-q}\,\mathrm{d}q\,\mathrm{d}x \tag{S.16}$$

$$= 0\,, \tag{S.17}$$

which will hold if

$$v\frac{\mathrm{d}U}{\mathrm{d}x} = \Lambda(x,v) - \Lambda(x,-v)\,. \tag{S.18}$$

Then from (2), we have

$$\Lambda(x,v) = \left(0 \vee v\frac{\mathrm{d}U}{\mathrm{d}x} \right) + \Gamma(x,v)\,. \tag{S.19}$$

Substituting (S.19) into (S.18), we see that (S.18) is satisfied and hence (S.16) is satisfied, meaning that we have demonstrated Equation (S.11) as required.

□

The SeZZ process forms the basis for NuZZ, which uses numerical approximations to find the roots of (S.1). Since these approximations are dependent on the last switching location through the generation of quadrature rules, NuZZ is not Markov. However expansion of the state space to include the last switching point does induce a Markov process, but the calculations that this gives rise to have proven unwieldy, hence the approach taken in this work.

S.2 House prices data

In this section we compare our algorithms on real-world data, specifically on the House-Price dataset from Kaggle [2017].

Consider a Bayesian linear regression model, $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We are interested in finding the posterior distribution over the length-d parameter vector $\boldsymbol{\beta}$, given the $n \times d$ matrix of independent variables $\boldsymbol{X}$, and the length-n vector of dependent variables, $\mathbf{y}$. The errors in the length-n vector $\boldsymbol{\varepsilon}$ are taken to be normally distributed, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$, where $\mathbf{0}$ is a length-n vector with all zero elements and $\boldsymbol{I}$ is the $n \times n$ identity matrix. Taking the prior on $\boldsymbol{\beta}$ to be improper, i.e. $\pi_0(\boldsymbol{\beta}) \propto 1$, the marginal posterior on the parameter vector given the observed data and standard deviation of the error has the simple form $\boldsymbol{\beta}|\{X, \mathbf{y}, \sigma\} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$, where we use hats to denote maximum-likelihood estimates (MLEs).

The House-Prices dataset is composed of $n = 1459$ house sales prices, along with $d = 80$ variables for each house, describing features such as position, type of heating, and type of bath. We selected this dataset because

it has a reasonably high number of variables, many of which are categorical, as is often the case in real datasets. Experiments performed on synthetic datasets, other simple targets and different performance metrics suggest that these conclusion hold in more general cases. Missing values in the data were dealt with as suggested in Gaudreau [2017], and we coded categorical variables using dummy variables. We discarded the variables *Utilities* (type of utilities available), *TotalBsmtSF* (total square feet of basement area), and *GrLivArea* (above grade (ground) living area square feet), to avoid singularities in the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and the variable *Id* (identity number), which is not part of the analysis. Finally, we make the standard choice to use the logarithm of the sales price as the variable $\mathbf{y}$. This processing step left us with $d = 77$ variables.

To simplify running the algorithms, we do not estimate the error variance parameter σ^2 as part of the MCMC algorithm. Rather, we set it exactly equal to its MLE, $\sigma = \hat{\sigma}$. This model has the advantage, therefore, of having known posterior, which allows us to use the same performance metric that we used in the previous tests.

The House Prices dataset is quite challenging for all the algorithms involved in this test. Its posterior combines multiple features that we previously explored separately. The number of variables is large ($d = 77$), their marginals have very different scales with a difference up to several orders of magnitude, and some of the variables are significantly correlated. All of these features commonly occur in real world datasets, which further increases the importance of this test. Due to the challenging nature of this model, the computational budget for this test was increased to 2.4×10^7 epochs, thinned down to 6×10^6 points, and the algorithms were started at the Maximum Likelihood Estimate. The results can be seen in Figure S.1.

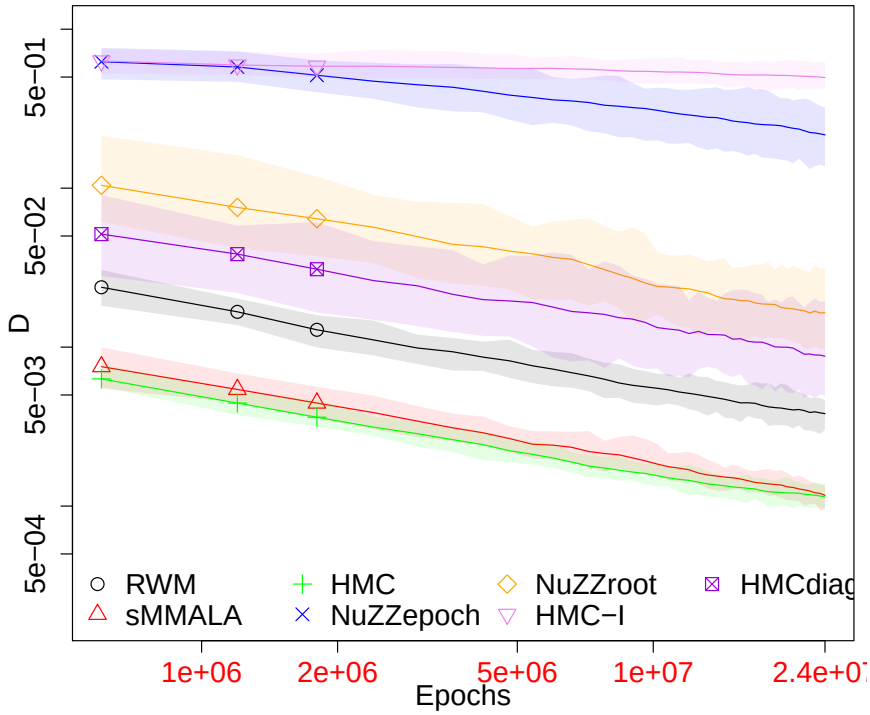

Figure S.1: Convergence of MCMC algorithms in the Kolmogorov-Smirnov metric on a linear regression model estimated on the House Prices datase.

sMMALA performs well, surpassing RW by good measure and performs similarly to HMC, again due to the low number of leapfrog steps that HMC needs, and to the fact that the Hessian is independent of $\boldsymbol{\beta}$.

Once again the green + line representing HMC with mass matrix $M = \hat{\sigma}^{-2}(\boldsymbol{X}^\top \boldsymbol{X})$, step size .83 and 3 leapfrog steps, performs very well, with acceptance close to 67%.

The pink line $\triangledown$ at the top of the graph on the other hand, shows the performance of HMC should the mass matrix be left as the identity matrix. That requires HMC to take an extreme 2400 leapfrog steps with step size .2, with poor results compared to the other algorithms. This may serve as a caveat to practitioners on the importance of tuning for HMC.

As in the other examples, the blue line $\times$ corresponding to NuZZepoch is far above the other algorithms, as the cost of the root finding and integration is very high.

The yellow $\Diamond$ line corresponding to NuZZroot is above the black $\bigcirc$ line corresponding to the Random Walk. The Zig-Zag dynamics in this example is hampered by the presence of features such as high dimension and

high correlation together. For comparison, we added the purple line ☒ representing HMC run with 100 leapfrog steps, step size .2, and mass matrix equal to the diagonal of $\hat{\sigma}^{-2}(\boldsymbol{X}^\top \boldsymbol{X})$, i.e. eliminating information about the correlation from the HMC setup, to make it more comparable to how we tune the velocities for NuZZ. The NuZZroot performance is still inferior to that of HMC in this case, likely due to the presence of correlation between regressors, once again highlighting the importance of the work in Bertazzi and Bierkens [2020].

S.3 Effects of tolerances

As mentioned in the text, in most experiments we set both the integration routine absolute tolerance and Brent's method absolute tolerance to 10^{-10}.

In the example from Section 5.7 we changes these tolerances systematically and assessed their impact on accuracy of the posterior samples.

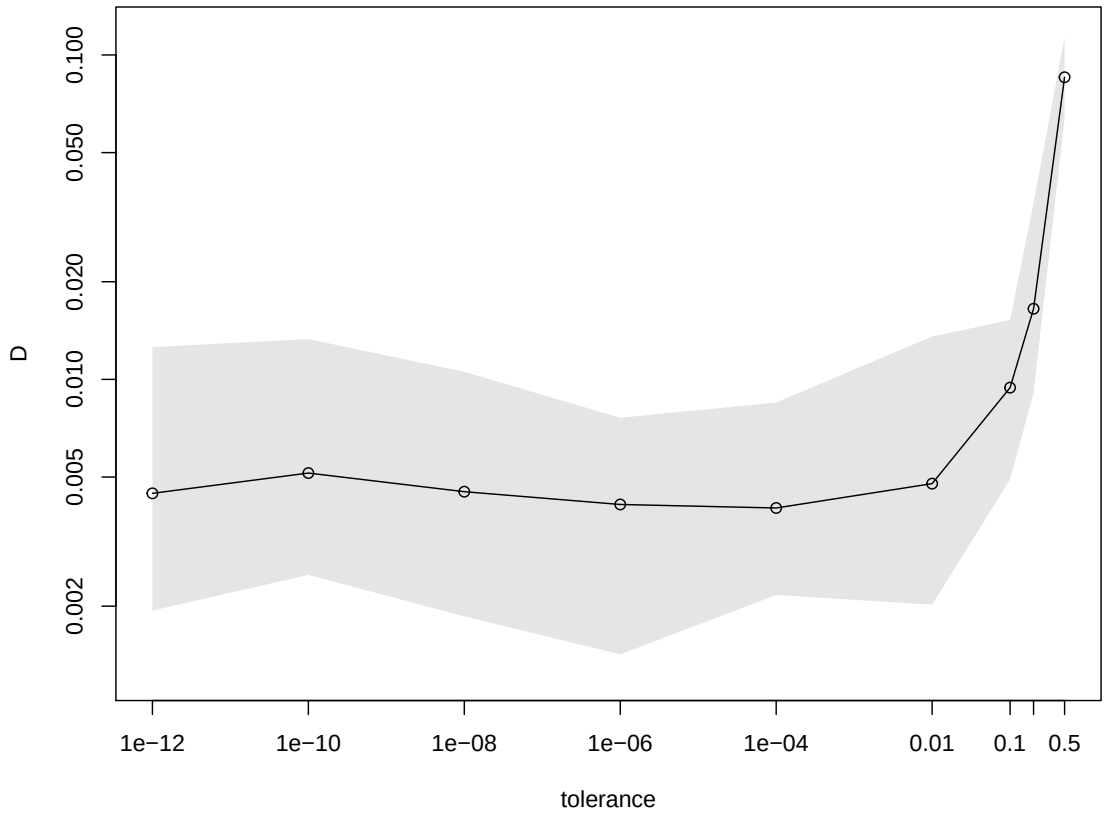

Figure S.2: Dependence between the largest Kolmogorov distance on the marginals and the root finding tolerance ε_{Bre}, for a fixed integration tolerance $\varepsilon_{\text{int}} = 10^{-10}$. The computational budget was 6×10^6 epochs, which was processed into 6×10^6 NuZZ samples.

Figure S.2 shows how D, the worst Kolmogorov-Smirnov distance on the marginals, increases as we increase ε_{Bre}. In this example, with 6×10^6 samples, the Monte Carlo error dominates the numerical error for all $\varepsilon_{\text{Bre}} < 10^{-2}$. This result is model dependent, but it supports the point that even though NuZZ uses numerical approximations to sample the switching times, if the numerical error is small, the difference in the posterior is not detectable.

We also tested the influence of the integration error on the quality of the sample from NuZZ, but since the QAGS integration routine performs a minimum of a 15-point Gaussian integration steps, this is often more than enough to reach the condition $\eta_{\text{int}} \leq \varepsilon_{\text{int}} = 10^{-10}$. Therefore we were unable to coarsen this part of the approximation sufficiently in order to produce the analogous plot to Figure S.2 for ε_{int}, but simply note that for all the examples we considered, one integration step is usually sufficient, and the integration tolerance simply acts as a fail-safe for when they are not.